

Das Führungskompetenzmodell des "Adaptive Leadership Competency Profile (ALCP)": Eine Gültigkeitsprüfung mithilfe der Probabilistischen Testtheorie

Bluemke, Matthias; Heene, Moritz; Bipp, Tanja; Steinmayr, Ricarda

Postprint / Postprint

Zeitschriftenartikel / journal article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:

GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:

Bluemke, M., Heene, M., Bipp, T., & Steinmayr, R. (2014). Das Führungskompetenzmodell des "Adaptive Leadership Competency Profile (ALCP)": Eine Gültigkeitsprüfung mithilfe der Probabilistischen Testtheorie. *Wirtschaftspsychologie*, 1, 81-94. <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-70080-9>

Nutzungsbedingungen:

Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.

Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use:

This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

By using this particular document, you accept the above-stated conditions of use.

Das Führungskompetenzmodell des „Adaptive Leadership Competency Profile (ALCP)“: Eine Gültigkeitsprüfung mithilfe der Probabilistischen Testtheorie

Matthias Bluemke¹, Moritz Heene², Tanja Bipp³ & Ricarda Steinmayr⁴

¹ Universität Heidelberg, Deutschland

² Ludwig-Maximilians-Universität München, Deutschland

³ Eindhoven University of Technology & Open University of the Netherlands, Niederlande

⁴ Technische Universität Dortmund, Deutschland

Zusammenfassung

Ziel des Artikels ist die Analyse der psychometrischen Güte der deutschen Version des „Adaptive Leadership Competency Profile“ (ALCP) (Sherron, 2000), eines Fragebogens zur Erfassung elf globaler Führungskompetenzen. Vorgestellt wird ein probabilistisch-testtheoretisches Vorgehen zur Überprüfung von mit (Führungs-)Kompetenzmodellen verbundenen Messmodellen auf der Grundlage des Rasch-Modells. Mit Hilfe des ordinalen Rasch-Modells wurden Items und Skalen analysiert und Rasch-Homogenität überprüft (N = 311). Mit Hilfe des Mixed-Rasch-Modells wurde die Annahme der gemeinsamen Skalierbarkeit von unterschiedlichen Mitarbeiterinnen und Mitarbeitern analysiert. Ein 11-dimensionales Rasch-Modell explorierte die latenten Korrelationen der Personenparameter. Sämtliche Skalen trennten Führungskräfte und Nicht-Führungskräfte nach Kompetenzgraden. Stärken des Ansatzes und praktische Implikationen werden diskutiert.

Schlüsselwörter: Führungskompetenzen, IRT, probabilistische Testtheorie, (Mixed) Rasch-Modell

The ‘Adaptive Leadership Competency Profile (ALCP)’: Validation on the basis of probabilistic test theory

Abstract

We analyze the psychometric properties of the German version of the ‘Adaptive Leadership Competency Profile’ (ALCP) (Sherron, 2000), which contains eleven scales assessing broad leadership competencies. We illustrate an approach to the probabilistic test-theoretical analysis of competency models according to the Rasch model. Based on the ordinal Rasch model, item and scale analyses demonstrated Rasch homogeneity (N = 311). A mixed Rasch model was used to inspect whether the different types of staff can be scaled jointly. An 11-dimensional Rasch model was used to explore the latent correlations among the person parameters. All scale means discriminated between executive managers and staff members. We discuss strengths of the approach and practical implications.

Key words: Leadership competencies, IRT, probabilistic test theory, (mixed) Rasch model

Einleitung

In den letzten Jahren haben Kompetenzmodelle der Führung weltweit zunehmend Eingang in die betriebliche Diagnostik des Entwicklungsstandes gegenwärtiger und zukünftiger Führungskräfte gefunden (Harris, 1998). Häufig entwickeln Unternehmen ihre eigenen Kompetenzmodelle, doch deren Reichweite erscheint gering und ihr Nutzen sowie die psychometrischen Eigenschaften ihrer Operationalisierungen bleiben oftmals unklar. Unabhängig vom theoretischen Status oder empirischen Mehrwert dieses Ansatzes zeigen wir beispielhaft, wie schrittweise mittels Item-Response-Theorie

die Qualität von Instrumenten zur Kompetenzerfassung überprüft werden kann, so dass deren Einsätze psychometrisch gerechtfertigt erscheinen. Wir überprüfen modellhaft die Anwendbarkeit des ordinalen Rasch-Modells für Skalen eines Führungskompetenz-Fragebogens – der deutschen Version des „Adaptive Leadership Competency Profile“ (ALCP) (Sherron, 2000). Anhand des Mixed-Rasch-Modells testen wir exploratorisch, ob die Fragebogenskalen für die analysierten Personen homogen sind, d.h. ob zu Recht angenommen wird, dass bei allen Personen die gleichen Kompetenzen gemessen wurden, oder aber ob Subgruppen identifizierbar sind, die sich aufgrund anderer als der intendierten Eigenschaften systematisch in ihrer Skalenbearbeitung unterscheiden. Wir beleuchten die Skalenstruktur anhand der Interkorrelationen sowie der latenten Korrelationen der Personenparameter mit Hilfe eines mehrdimensionalen Rasch-Modells.

Theoretischer Hintergrund: Kompetenzen und Kompetenzmodelle

Mit [berufsrelevanten] Kompetenzen werden für eine *erfolgreiche Zielerreichung* [im Unternehmen] wichtige Verhaltensweisen zusammengefasst, die relativ überdauernden Eigenschaften ähneln, aber als durch Selbstorganisation (wie berufliche Weiterbildungsmaßnahmen oder Berufserfahrung) veränderbare Größen aufgefasst werden (Erpenbeck & Rosenstiel, 2003). Kompetenzen werden von anderen psychologischen Merkmalen durch ihren *expliziten Leistungs- bzw. Erfolgsbezug* abgegrenzt (McClelland, 1973). Als Erfassungsmethode werden aber wie bei den anderen Ansätzen häufig (Selbst-)Ratings eingesetzt. Erfolgsbezug sollte nicht nur, aber auch, auf der semantischen Ebene der Items hergestellt werden. Anders als bei anderen Eigenschaften, Verhaltensweisen und Kognitionen wird erwartet, dass er durch Korrelationen mit Erfolgskriterien (Performanz, Gehalt) oder durch die Trennung erfolgreicher von nicht-erfolgreichen Personen nachzuweisen ist (Catano, 1998; Popper, Amit, Gal et al., 2004).

In Kompetenzmodellen werden Einzelkompetenzen in hierarchischen Katalogen gruppiert, was eine gemeinsame Sprache im Personalbereich hinsichtlich Fach-, Sozial- und Personalkompetenzen sowie eine *situationsspezifische Auswahl* der als erfolgsrelevant erachteten Merkmale gestattet (Lager & O'Connor, 2000). In Unternehmen werden erfolgskritische Verhaltensweisen aus mehreren Quellen (z.B. Vorgesetzten oder Kollegen) abgeleitet (Brannick & Levine, 2002). Untermuert werden diese Modelle durch firmenspezifische Anforderungsanalysen, die unternehmensspezifische Verhaltenserwartungen an Führungskräfte als Kriterien nutzen (Schuler, 2001; Rosenstiel, 2003).

Werden Kompetenzmodelle, die auf Führung abzielen, auch zur Identifizierung von Nachwuchsführungskräften genutzt, sollen sie gleichzeitig aber auch bei Nachwuchskräften, die keine *Führungserfahrung* aufweisen, *Führungspotential* identifizieren. Ob bereits ein ausreichendes Kompetenzniveau vorliegt, sollte deshalb unabhängig von aktuellen beruflichen Positionen messbar sein. Zur objektiven Kompetenzmessung eignen sich Leistungstests und Simulationsübungen (Assessment Center). Aus Gründen der Ökonomie stellen auch Selbstberichte *eine* Methode der Wahl dar, denn nicht immer liegen bei unternehmensexternen Kandidaten oder Trainees objektive Verhaltensdaten oder Fremdbeurteilungen vor. Jedoch unterliegen selbst verhaltensnahe Selbstberichte als subjektive Auskünfte klar der Verfälschbarkeit. In keinem Fall sollten Items eine aktuelle Führungsrolle oder Führungserfahrung des Befragten implizieren oder gar explizit für die Kompetenzerfassung voraussetzen. Ein Blick auf Standardinstrumente zum Führungsverhalten wie den Leader Behavior Description Questionnaire (Stogdill, 1963) oder den Multifactor-Leadership-Questionnaire (Felfe, 2006) zeigt aber, dass Instrumente teilweise eine disziplinarische bzw. supervisierende Führungsrolle voraussetzen.

Selbst bei Vermeidung solcher Fehlerquellen stellt sich die Frage, inwieweit sich die Kompetenzen von Personen mit unterschiedlichen Tätigkeiten bzw. Berufshintergründen mit demselben Instrument erfassen und vergleichen lassen. So könnten zwei (oder gar mehr) Untergruppen vorliegen, deren unterschiedliches Verständnis der Kompetenzitems bzw. deren abweichende Skalenverwendung eine Vergleichbarkeit der Ergebnisse beeinträchtigt. Dies könnte z.B. eine Kompetenzmessung bei führungserfahrenen vs. -unerfahrenen Personen (oder auch Männer und Frauen) mit Hilfe derselben Items und Skalen gefährden.

Probleme von Kompetenzerfassungen

Manche der im Folgenden skizzierten Probleme tangieren auch andere diagnostische Zugänge zur Kompetenzerfassung (z.B. verhaltensnahes Assessment Center; 360°-Feedback); im Mittelpunkt stehen bei uns jedoch Selbstberichtverfahren, analog zu klassischen Persönlichkeitsverfahren. Die Schwierigkeiten der gängigen Praxis der Kompetenzmessung sind evident: Trotz wiederholter Forderung nach stärkerem Theoriebezug bei der Vorhersage von Berufs- bzw. Führungserfolg (z.B. Tett, Jackson, Rothstein & Reddon, 1999) lassen Heterogenität der Kompetenzmodelle und fehlender Theoriebezug kaum äquivalente Operationalisierungen oder gar Replizierbarkeit ihrer Vorhersagegüte zu. Obwohl die akkurate Messung und Beurteilung von Personmerkmalen im Interesse von Diagnostikern und Unternehmen liegt und obwohl die mit wissenschaftlichen Methoden überprüfte Brauchbarkeit zu den essentiellen Anforderungen

an psychodiagnostische Verfahren gehört (Testkuratorium, 1986), werden psychometrische Gütekriterien von Instrumenten zur Kompetenzerfassung entweder kaum überprüft oder bleiben dem Anwender regelmäßig unbekannt (Schippmann, Ash, Battista, Carr, Eyde, Hesketh et al., 2000).

Werden in der Praxis unterschiedliche Item- und Antwortformate zur Kompetenzmessung kombiniert, erfordert dies eine Modelltestung, ob ihre Aggregation zu Skalensummenscores als Kompetenzindikatoren gerechtfertigt ist. Wird eine Kompetenz anhand einer Auswahl von inhaltlich passenden Items erhoben, muss die Itemauswahl die Schätzung des Kompetenzgrades auf möglichst qualitativ hohem Niveau ermöglichen. Werden einzelne Skalen eines Instruments situationsspezifisch ausgewählt, dann sollten diese Skalen ebenfalls psychometrisch belastbar sein. Sofern Kompetenzmodelle berufs- oder länderübergreifend eingesetzt werden, tritt die Anforderung hinzu, Kompetenzen messinvariant, also qualitativ gleichwertig, unabhängig von Ort, Alter, Geschlecht, fachlichem Hintergrund, Sprache und (Organisations-)Kultur zu erfassen.

Vorteile einer probabilistischen Fundierung von (Führungs-)Kompetenzmessungen

Im Unterschied zur herkömmlichen Entwicklung von Tests auf Basis der klassischen Testtheorie (KTT) stellt die Item-Response-Theorie (IRT) Modelle zur Verfügung, welche die psychometrischen Eigenschaften von Instrumenten zur Erfassung von beruflichen (und damit auch führungsspezifischen) Kompetenzen empirisch überprüfbar machen. Hierzu gehört in erster Linie,

- a) dass alle Items einer Skala dasselbe Personmerkmal erfassen (Itemhomogenität) und
- b) dass alle Personen die Skalen aufgrund desselben Personmerkmals beantworten (Personenhomogenität).

Darüber hinaus bietet die IRT den Vorteil, dass die verwendeten Antwortformate der Ratingskalen (z.B. Anzahl der Antwortkategorien oder Kompetenzstufen) hinsichtlich ihrer psychometrischen Güte überprüft und bei Bedarf angepasst werden können. Dies sind große Vorteile gegenüber der korrelationsbasierten klassischen Testtheorie, die für diese Probleme keine Methoden bereithält. Der IRT-Ansatz weist somit bei der Kompetenzmessung mehrere Vorzüge auf (vgl. Rost, 2004; für ausführlichen Vergleich zwischen KTT und IRT siehe bspw. Bond & Fox, 2007).

Die job- und firmenübergreifende Anwendbarkeit von Instrumenten zur Kompetenzmessung kann auf Grundlage des Rasch-Modells geprüft werden. Kompetenzmessung mit einer Raschmodellierung ermöglicht spezifisch objektive Vergleiche sowie eine Skalierung der Kompetenzen und Testpersonen auf derselben Skala (sog. Logit-Skala). Spezifisch objektive Vergleiche implizieren bei Vorliegen der Modellgeltung, dass Differenzen in den Kompetenzgraden zwischen Testpersonen unabhängig von den verwendeten Items – und Differenzen zwischen Itemschwierigkeiten unabhängig von der Verteilung der Kompetenzausprägungen – interpretiert werden können. Wenn diese Voraussetzungen erfüllt sind, können Personen mit aktuell unterschiedlichen Tätigkeiten bzw. Berufshintergründen gemeinsam skaliert und verglichen werden.

Die Annahme von stichprobenunabhängigen Itemparametern (Itemschwierigkeiten) stellt sich empirisch oft als nicht haltbar heraus. Eine Abschwächung der Annahme besteht darin anzunehmen, dass das Rasch-Modell nicht in der gesamten interessierenden Population gilt, sondern jeweils nur innerhalb von Personengruppen („Klassen“). Diesen Ansatz hat Rost (1990) mit der Entwicklung des Mixed-Rasch-Modells verfolgt. Das Mixed-Rasch-Modell kann auch als strenges Verfahren zur Überprüfung der Personenhomogenitätsannahme herangezogen werden, da es nach latenten Klassen sucht, die sich in ihren Antwortmustern maximal unterscheiden und hierüber definierbaren Gruppen zugeordnet werden können. Dies gestattet eine Abschätzung, ob Items heterogen interpretiert bzw. Antwortskalen heterogen genutzt werden. Ist dies nicht der Fall, können alle Personen gemeinsam skaliert und hinsichtlich ihrer Kompetenzgrade verglichen werden, also auch Personen unterschiedlicher Profession oder Hierarchie.

Testkonstruktion – Gütekriterien

Der nachfolgend beschriebene Weg einer Testkonstruktion dient als Leitmodell für die Entwicklung firmenspezifischer oder -übergreifender Instrumente zur Kompetenzerfassung. Wir tragen damit Markus, Cooper-Thomas und Allpress (2005) Rechnung, die auf den „urgent need to improve the validity of the competency models in use“ (S. 124) hinwiesen. Zunächst soll auf die Aspekte Itemhomogenität, Schwellenparameter und Personenhomogenität genauer eingegangen werden, da diese einer Überprüfung standhalten müssen, um zu zeigen, dass die im Rasch-Modell geltenden Annahmen für das Kompetenzverfahren tatsächlich gültig sind. An dieser Stelle ist wichtig, dass wir in dieser Studie aufgrund des mehrstufigen Antwortformates der Ratingskalen das *ordinale* Rasch-Modell (auch Partial-Credit-Modell genannt) angewandt haben, das eine Erweiterung des dichotomen Rasch-Modells darstellt (Masters, 1982).

Itemhomogenität

Wie gut ein einzelnes Item zu den Annahmen des ordinalen Rasch-Modells passt, beschreiben Item-fit-Indices. Passen die beobachteten Itemantwortmuster zu den Antwortmustern, die man unter der

Geltung des Rasch-Modells annimmt, kann von einer homogenen Erfassung der Merkmalsdimension ausgegangen werden. Dabei ist neben einer zu schlechten Itempassung (Underfit) auch eine zu gute Itempassung (Overfit) durch deterministische Antwortmuster denkbar. Ein derartiger Overfit kann ein Anzeichen von Itemredundanzen, Itemmehrdimensionalität oder Iteminteraktionen sein (z.B. Tuerlinckx & De Boeck, 2001). Underfit kann dadurch entstehen, dass die Antworten achtlos („zufällig“) gegeben werden oder aber durch missverständliche oder mehrdeutige Itemformulierungen zu Verständnisproblemen führen, die sich in inkonsistentem Antwortverhalten niederschlagen. Wie Smith (2005) zeigen konnte, führen sowohl Underfit als auch Overfit zu verzerrten, ungenaueren Schätzungen der Itemschwierigkeiten und Personenparameter.

Schwellenparameter

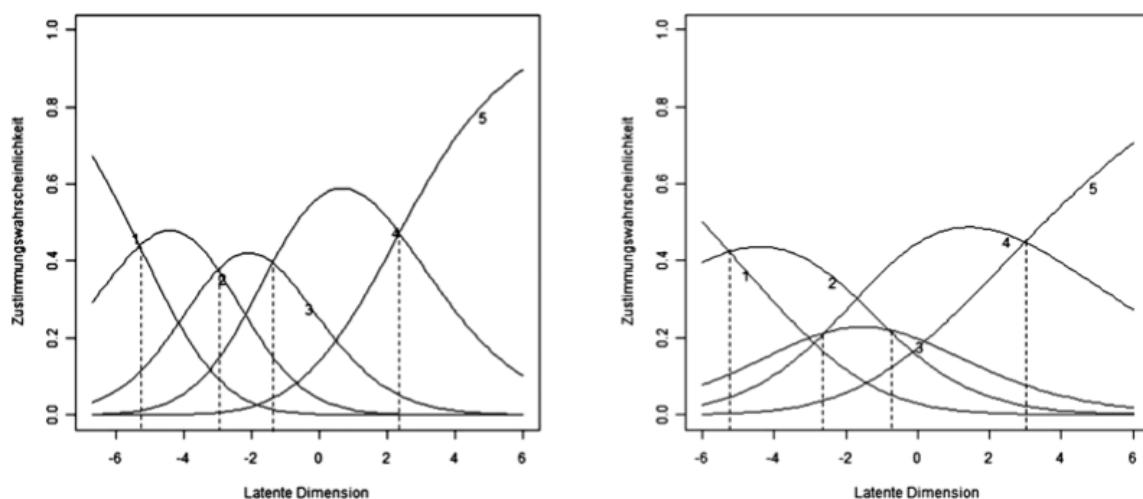
Anhand des ordinalen Rasch-Modells zeigen wir, wie analysiert werden kann, ob bei jedem Item die Antwortskalen wie intendiert verwendet wurden. Kennzeichnend sind hierbei die sogenannten Schwellenparameter zwischen *mehrfach* abgestuften Antwortkategorien. In Analogie zur Annahme der Stichprobeninvarianz der Itemparameter im dichotomen Rasch-Modell werden hier die Schwellenparameter als stichprobeninvariant angenommen. Die Schwellenparameter stellen dabei die Schnittpunkte benachbarter Kategorienfunktionen dar. Die Kategorienfunktionen beschreiben für jede einzelne Antwortkategorie die Wahrscheinlichkeiten, bei gegebener Personmerkmalsausprägung zu antworten.

Für eine sinnvolle Interpretation der Antwortskala als Kompetenzskala müssen für höherrangige Antwortkategorien höhere Kompetenzausprägungen nachweisbar, Schwellenparameter also entsprechend des Kompetenzzuwachses geordnet sein (s. Abbildung 1A für die Kategorienfunktionen eines Items mit geordneten Schwellenparametern). Bei ungeordneten Schwellenparametern (Abbildung 1B) kann bestimmten Antwortkategorien *kein* Abschnitt auf dem zugrundeliegenden Merkmal zugewiesen werden; die Antwortkategorien sind nicht aufsteigend geordnet und weisen kein Ordinalskalenniveau auf (Rost, 2004). In Abbildung 1B liegen die Schwellenparameter der Antwortkategorienfunktionen für Kategorie 2 und 3 *über* den Schwellenparametern der Antwortkategorienfunktionen 3 und 4. Die Ordinalskalengüte (Rangordnung der Antwortkategorien) ist verletzt, weil numerisch aufeinanderfolgenden Antwortkategorien keine aufeinanderfolgenden Abschnitte der latenten Dimension zugeordnet werden können. Schwellenparameter erweisen sich in der Praxis häufig dann als ungeordnet, wenn das Antwortformat zu differenziert ist, d.h. wenn es zu viele Antwortkategorien beinhaltet, die die Probanden nicht mehr hinreichend unterscheiden können. In solch einem Fall wäre eine Weiterentwicklung des Fragebogens anzudenken, wobei ein Antwortformat mit weniger Antwortkategorien verwendet würde.

Personenhomogenität

Zur Beantwortung der Frage, ob alle befragten Personen eine Kompetenzskala in gleicher Weise beantworten und die jeweilige Skala somit bei allen Personen dasselbe Merkmal erfasst, werden heterogene Personengruppen mit Hilfe des Mixed-Rasch-Modells ermittelt (Rost, 2004). Dieses Mischverteilungsmodell identifiziert Personenklassen, die sich in ihrem Antwortverhalten *maximal* unterscheiden. Vergleiche zwischen Modellen mit verschiedenen Klassenanzahlen anhand der deskriptiven Informationskriterien Bayesian Information Criterion (BIC) und Consistent Akaike Information Criterion (CAIC) gestatten die *relative* Modellpassung zu beurteilen, beispielsweise eine Zweiklassen- gegenüber einer Einklassenlösung, wobei niedrigere Werte für eine relativ bessere Modellpassung stehen. Beide Statistiken bevorzugen nach dem Parsimonieprinzip einfachere Klassenlösungen. Sind die untersuchten Personen hinsichtlich ihres Antwortverhaltens ho-

ABBILDUNG 1: Antwortkategorienfunktionen eines Items mit fünf Kategorien der Ratingskala: geordnete (Abb. 1A) bzw. ungeordnete (Abb. 1B) Schwellen



mogen, so müssen die Daten besser zu einer Ein- als zur Mehrklassenlösung passen.

Adaptive Leadership Competency Profile

Die bisherigen theoretischen Ausführungen werden veranschaulicht anhand des „Adaptive Leadership Competency Profile“ (ALCP) (Sherron, 2000) – einem Fragebogenverfahren zu einem Führungskompetenzmodell, das in einer umfangreichen Studie der North Texas Universität zu Führungskompetenzen in hochleistenden Organisationen für den unternehmensübergreifenden Einsatz auf Basis der IRT entwickelt wurde. Das Ziel war, einen zeitgemäßen Kriterienkatalog von intersubjektiv als praxisrelevant und erfolgskritisch eingestuften, globalen Führungskompetenzen zu erstellen. Bei der Itemgenerierung wurden qualitative mit deduktiven Strategien kombiniert (vgl. Bühner, 2006) sowie theoretische Anleihen bei verschiedenen Führungstheorien gemacht (bspw. situationale Führungstheorie, Kontingenztheorien der Führung, transformationale vs. transaktionale Führung). Basierend auf 600 Interviews mit aktiven Führungskräften und Mitarbeitern unterschiedlicher Unternehmen und Hierarchieebenen wurden 13 rollen- und statusunabhängige Verhaltensweisen abgeleitet, die nach inhaltlichen Überlegungen zunächst in 13 Kompetenzen gruppiert wurden (Spencer, McClelland & Spencer, 1990). Personen mit ausgewiesener Fachexpertise reorganisierten die 13 Kompetenzen nach Ähnlichkeit, die sodann elf Kompetenzskalen (mit insgesamt 65 Items) zugeordnet werden konnten: *Einfluss und Motivation, Lernen, Selbstmanagement, Visionäre Kompetenz, Teamentwicklung, Initiative, Ethisches Verhalten, Förderung von anderen, Kommunikation, Entscheidungsfindung und Veränderung* (vgl. Tabelle 1 für Beispiel-

TABELLE 1 Kompetenzskalen des ALCP, zugehörige Erläuterungen und Beispielitems

| Kompetenz | Items/Kompetenzfacetten |
|---|---|
| 1. Einfluss und Motivation | |
| Motivation und Beeinflussung durch Befriedigung grundlegender menschlicher Bedürfnisse, intellektuelle Stimulation oder individuelle Anreize | „Sorgt dafür, dass die Mitarbeiter/-innen ihr Bestes geben.“ |
| 2. Lernen | |
| Persönliche Realisation des Prinzips des lebenslangen Lernens und Unterstützung von Lernumgebungen, die selbständige Aufgabenbewältigung und Wissenserweiterung fördern | „Fördert die Idee des lebenslangen Lernens als Lebensmotto.“ |
| 3. Selbstmanagement | |
| Setzen von Leistungszielen, effektive Arbeitsorganisation und sinnvoller Einsatz von Ressourcen | „Nutzt Zeit und Ressourcen effizient.“ |
| 4. Visionäre Kompetenz | |
| Generieren eigener Visionen von zukünftigen Vorgängen oder Entwicklungen zu generieren | „Sieht Licht am Ende des Tunnels.“ |
| 5. Teamentwicklung | |
| Aktivitäten, die sowohl der Erreichung gemeinsamer Ziele dienen als auch die Kooperation und Zusammenarbeit fördern | „Unterstützt zur Erreichung gemeinsamer Ziele das Team.“ |
| 6. Initiative | |
| Entschlusskraft und die Fähigkeit, etwas zu beginnen bzw. anzuregen, wenn es die Situation erfordert | „Tut Dinge, bevor er/sie darum gebeten oder dazu durch äußere Umstände gezwungen wird.“ |
| 7. Ethisches Verhalten | |
| Ehrlichkeit, Integrität, Glaubwürdigkeit und Einsatz für in dem jeweiligen Kulturkreis geltende ethische Richtlinien | „Handelt nach Werten wie Wahrheit und Ehrlichkeit.“ |
| 8. Förderung von anderen | |
| Identifikation von Standards, die eine sinnvolle Beurteilung der Leistung anderer erlaubt sowie Potenzialerkennung bei Mitarbeiter/-innen | „Fördert die Fähigkeiten von Angestellten.“ |
| 9. Kommunikation | |
| Fundierte Feedback im Bedarfsfall und Vermittlung bestimmter Werte und Inhalte im Rahmen der Tätigkeit | „Spricht offen und direkt mit anderen über Leistungsschwächen.“ |
| 10. Entscheidungsfindung | |
| Treffen fundierter Entscheidungen auf Basis der Analyse von Prozessen und Problemen; Konsequenzen und Alternativen werden bedacht | „Trifft schwierige Entscheidungen und verfolgt sie dann weiter.“ |
| 11. Veränderung/Change-Management | |
| Veränderungsbereitschaft, Suche nach neuen Ideen und Unterstützung bereits in Gang gesetzter Veränderungsprozesse | „Betrachtet Veränderung als eine Chance zur Erneuerung.“ |

items). Einige Studien zum englischen Original des Fragebogens weisen die Rasch-Konformität der ALCP-Skalen nach (Sherron, 2002), wobei allerdings einige Items mit ungenügendem Modell-Fit nicht berücksichtigt wurden.

Das ALCP bildet eher globale und partiell überlappende Führungsaspekte ab und gilt in zweifacher Hinsicht als adaptiv: Erstens können im Unternehmen spezifisch relevante Kompetenzschwerpunkte gesetzt werden; es wird also ein modulares Kompetenzmodell verfolgt. Zweitens spielt der Name auf adaptive Führung an, welche die Bedeutung einzelner Personen, Rollen und Verhaltensweisen in der sich permanent wandelnden Unternehmensorganisation betont, um auf Probleme mit unklarer Lösung angemessen reagieren zu können, z.B. bei den Kompetenzen *Initiative* oder (lebenslangem) *Lernen*.

Die Konzeption des ALCP als *mehrdimensionales* Verfahren lässt die Anwendung elf eindimensionaler Rasch-Modelle wie oben beschrieben zunächst fraglich erscheinen. Allerdings existieren mehrdimensionale Verallgemeinerungen von Rasch-Modellen, welche die Itemantworten in Abhängigkeit von mehreren Personeneigenschaften modellieren können. Im vorliegenden Fall stellt das mehrdimensionale Item-Komponenten Rasch-Modell (Rost & Carstensen, 2002) ein für die weiteren Analysen zentrales Modell dar. Es ermöglicht, die Item-Skalenpassung aller elf Skalen *simultan* zu überprüfen und gleichzeitig Aufschluss über die messfehlerbereinigten Skaleninterkorrelationen zu erhalten. Da bei Kompetenzmodellen generell von hohen Skaleninterkorrelationen auszugehen ist, stellt dieser Ansatz das Mittel der Wahl dar.

Studie

Methodisches Vorgehen

Überprüft wurde die deutsche Übersetzung des ALCP (Sherron, 2000), das bereits eine Itemselektionsphase, wie sie auch aus der KTT bekannt ist, durchlaufen hat. Zugrunde legten wir eine eigene Übersetzung der vollständigen elf ALCP-Skalen basierend auf den originalen 65 Items. Wir untersuchten, ob die Kompetenzmessungen im deutschen Kulturkreis weiterhin Modellkonformität bzw. Rasch-Homogenität, also Skalen-Eindimensionalität und Messäquivalenz über verschiedene Personengruppen hinweg aufweisen, ob erwartete Kompetenzunterschiede unterschiedlich führungserfahrener Personen detektiert werden und wie die statistischen Skalenzusammenhänge ausfallen.

Methode

Durchführung und Stichprobe

Statt einer Stichprobe mit gemeinsamem Organisationsbezug wurden führungserfahrene und -unerfahrene Mitglieder eines deutschlandweiten Online-Panels zur Studie eingeladen. Teilnahmevoraussetzung war eine mindestens 3-monatige Berufstätigkeit in der aktuellen Position. Im Interesse einer ausreichenden Heterogenität wurden Personen mit und ohne aktuelle Führungsposition zu Vergleichszwecken zugelassen, um die Messung unterschiedlicher Kompetenzniveaus mit Hilfe derselben Skalierung sicherstellen zu können.

Während der dreiwöchigen Umfragedauer wurde der Webfragebogen 1.055-mal aufgerufen; 371 Personen machten Teilangaben. Ausgeschlossen aus der Analyse wurden Personen, die vor dem Ende abbrachen (verbleibendes $N = 311$; Unit-non-response-Rate = 65%; Drop-out-Rate = 16%; Beendigungsquote = 30%). Fehlende Antworten wurden skalenweise mittels des iterativen EM-Algorithmus geschätzt (Schafer & Graham, 2002). Die resultierende Stichprobe von 311 Probanden war geschlechterbalanciert (51,4% Männer; Altersspanne = 19-67 Jahre, $M = 39.4$, $SD = 10.5$). Sie stammten aus einem breiten Branchenspektrum (Automobil, Banken/Versicherung, Bau, Bildung & Wissenschaft, chemische Industrie, EDV & IT, Energie, Gesundheitswesen, Lebensmittel, Luftfahrt, Marketing, Medizin, Öffentlicher Dienst, Polizei, Sozialwesen, Transport, Unternehmensberatung). Die Mehrheit (63%) verfügte über Schulabschlüsse auf dem Niveau von mindestens (Fach-)Hochschulreife, überwiegend auch über Führungserfahrungen (69.6%). Zum Zeitpunkt der Erhebung bekleideten aktuell 36.7% innerhalb ihrer Organisationen eine Position mindestens auf Gruppen- oder Abteilungsleitungsebene (Geschäftsführung: 21, [Erweiterter] Vorstand: 7, Sektions-/Spartenleitung: 8, Abteilungsleitung: 31, Gruppenleitung: 49, Stabstelle: 23, Sonstige: 172). Im Mittel waren diese Personen seit knapp sieben Jahren Führungskraft ($M = 6.92$; $SD = 6.67$) mit einer Führungsspanne von 24.65 ($SD = 91.74$; Median = 6) zugeordneten Personen.

Materialien

Die ALCP-Items wurden zunächst ins Deutsche übersetzt und dann von einer bilingualen Englischlehrerin ins Englische rückübersetzt. Geringfügige Abweichungen führten nach Rücksprache zu leichten Modifikationen der Wortlaute. Die Kompetenzfacetten wurden von den Versuchspersonen auf einer Likert Skala beurteilt, ob die gemachten Aussagen auch auf sie persönlich zutreffen (1 = „völlig unzutreffend“ bis 5 = „sehr zutreffend“).

Ergebnisse

Modellkonformität und Skalenkennwerte

Die folgenden Analysen wurden mit Hilfe der Programme WINMIRA (Davies, 2001) und Acer Con-Quest 2.0 (Wu, Adams, Wilson & Haldane, 2007)

TABELLE 2 Kennwerte der ALCP-Items auf Skalenebene

| | Anzahl Items | Q-Statistik Itemebene | | z(Q) | | Modellfit | | | |
|---|--------------|-----------------------|-----|------|------|------------------|--------|-------------------|--------|
| | | Min | Max | Min | Max | Einklassenlösung | | Zweiklassenlösung | |
| | | | | | | BIC | CAIC | BIC | CAIC |
| Einfluss und Motivation | 6 | .06 | .20 | -.82 | 1.29 | 4035.1 | 4060.1 | 4079.3 | 4130.3 |
| Lernen | 5 | .04 | .10 | -.52 | .49 | 3412.5 | 3433.5 | 3432.2 | 3475.2 |
| Selbstmanagement | 8 | .06 | .13 | -.67 | .63 | 4909.1 | 4942.1 | 4963.3 | 5030.3 |
| Visionäre Kompetenz | 5 | .07 | .11 | -.45 | .33 | 3821.7 | 3842.7 | 3784.9 | 3827.9 |
| Teamentwicklung | 7 | .05 | .15 | -.86 | 1.16 | 4560.9 | 4589.9 | 4600.2 | 4659.2 |
| Initiative | 3 | .05 | .09 | -.47 | .33 | 2118.2 | 2131.2 | 2146.5 | 2173.5 |
| Ethisches Verhalten | 6 | .03 | .08 | -.52 | .48 | 3877.4 | 3902.3 | 3911.0 | 3962.0 |
| Förderung von anderen | 6 | .05 | .08 | -.57 | .56 | 4045.1 | 4070.1 | 4090.6 | 4141.6 |
| Kommunikation | 5 | .06 | .13 | -.43 | .79 | 3404.3 | 3425.3 | 3415.7 | 3458.7 |
| Entscheidungsfindung | 7 | .08 | .12 | -.41 | .38 | 4811.1 | 4840.1 | 4913.6 | 4972.6 |
| Veränderung | 7 | .08 | .13 | -.66 | .56 | 4782.1 | 4811.1 | 4822.8 | 4881.8 |
| Anmerkungen: Q-Statistiken aus Einklassenlösung des Mixed-Rasch-Modells. | | | | | | | | | |

durchgeführt. Die Antworten wurden separat für die elf Einzelskalen anhand des ordinalen Rasch-Modells mit WINMIRA modelliert (siehe Tabelle 2 für eine Zusammenfassung der Itemkennwerte).

Itemhomogenität

Im Fall der Software WINMIRA wird zur Beurteilung der Items der Q-Index herangezogen (Rost & von Davier, 1994). Er beschreibt die Likelihood eines beobachteten Itemantwortmusters unter den Modellannahmen und kann Werte zwischen 0 (deterministisches Antwortmuster) und 1 (minimale Likelihood unter den Modellannahmen; ggf. maximal negative Itemtrennschärfe bei negativ korrelierenden Items) annehmen. Rost (2004) spezifiziert für den Q-Index Werte zwischen .10 und .30 als akzeptable Itempassung, wobei kein Item signifikant von seinem Erwartungswert unter den Modellannahmen mit ($|z(Q)| > 1.96$) abweichen darf. Laut Q-Statistiken wichen die beobachteten ALCP-Itemantworten nicht signifikant von den unter Modellgeltung zu erwartenden Antwortmustern ab. Damit zeigten sämtliche Items Modellpassung. Die Itemhomogenität kann angenommen werden. Für jede einzelne ALCP-Skala kann die Annahme der Rasch-Modellkonformität zunächst beibehalten werden.

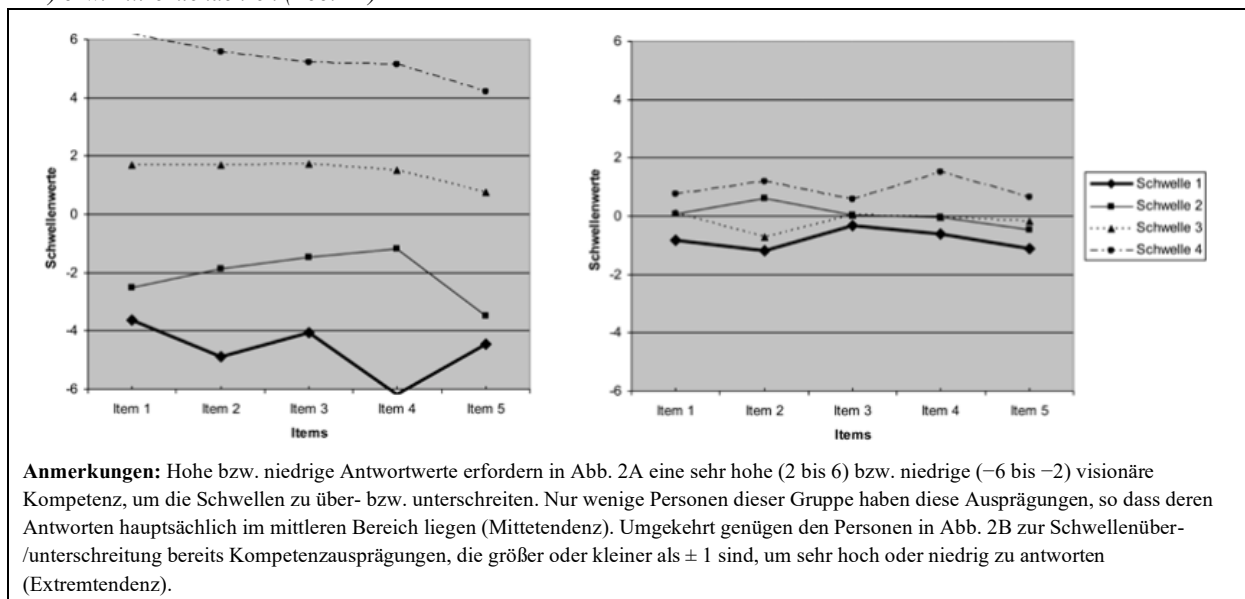
Schwellenparameter

Eine Inspektion der Schwellenparameter der 65 ALCP-Items zeigte, dass die Antwortkategorien im Wesentlichen aufsteigend geordnet sind, mit folgenden, für das vorliegende Itemformat bestehenden Qualitätseinschränkungen: Bei elf Items der Skalen *Selbstmanagement*, *Visionäre Kompetenz*, *Teamentwicklung*, *Initiative*, *Förderung von anderen*, *Kommunikation* und *Veränderungen* zeigten sich Vertauschungen der Schwellenparameter der ersten und zweiten Antwortkategorie („völlig unzutreffend“ – „unzutreffend“). Die unterste Antwortkategorie wurde seltener gewählt als aufgrund der Verteilung der Kompetenzausprägungen zu erwarten war. Beispielsweise entschieden sich bei den acht *Selbstmanagement*-Items durchschnittlich nur 0,6% bzw. 2,7% der Stichprobe für „völlig unzutreffend“ bzw. „unzutreffend“. Eine Möglichkeit wäre, diese Antwortkategorien für Itemanalysen zusammenzufassen; um aber die numerische Bedeutung der Antwortoptionen und eine hohe Vergleichbarkeit zum Originalfragebogen zu erhalten, verzichteten wir auf die Zusammenlegung. Für die übrigen Items aber ist empirisch anzunehmen, dass mit zunehmender Zustimmung zu den einzelnen Items insgesamt ein größerer Kompetenzgrad der Person einhergeht.

Personenhomogenität

Als nächstes wurde die Personenhomogenität mit einem Mixed-Rasch-Modell geprüft. Der Modellfit der Ein- und Zweiklassenlösung wurde anhand der Informationskriterien *BIC* und *CAIC* vorgenommen (Rost, 2004). Das Minimum der beiden Stichprobengrößen-korrigierten Informationskriterien weist auf das bessere Verhältnis von Anpassungsgüte und Modellparameterzahl hin. Obwohl die relativ kleine Stichprobengröße keine zuverlässige Schätzung der Itemparameter der Zweiklassenlösung erlaubt, sind die Abstände zwischen den Einklassen- und Zweiklassenmodellen deutlich (Tabelle

ABBILDUNG 2 Schwellenparameter der Skala Visionäre Kompetenz in den beiden Personenklassen mit Mittetendenz (Abb. 2A) bzw. Extremtendenzen (Abb. 2B)



2). Der Schluss liegt nahe, dass Einklassenmodelle die Daten (mit Ausnahme der Skala *Visionäre Kompetenz*) angemessener beschreiben. Wo die Einklassenlösung den niedrigeren Kennwert erreicht, ist die Einführung einer weiteren Klasse statistisch nicht angebracht. Sollten Zweiklassenlösungen die Daten besser als Einklassenlösungen beschreiben, empfiehlt sich eine Ursachenforschung (Kreuztabellierung) anhand verschiedener Variablen (bspw. Geschlecht oder Führungserfahrung). Allerdings *kann* die Zweiklassenlösung mit einer bekannten Klassifizierungsvariable wie Geschlecht oder Führungserfahrung korrespondieren, muss es aber nicht. Die möglicherweise vorhandenen zwei Klassen können sich auch durch zunächst unbekannte Klassenmerkmale ergeben.

Bei der Skala *Visionäre Kompetenz* ist – sowohl nach *BIC* als auch nach *CAIC* – die Zweiklassenlösung passender. Ein Vergleich der Schwellenparameter der beiden Gruppen legte nahe, dass Antworttendenzen der Befragten für die latenten Klassen verantwortlich sind (Majj-de Majj, Kelderman & van der Flier, 2008). Die erste Klasse (64%) bildeten Personen mit einer Tendenz zu mittleren Antwortkategorien. Für sie war die erste Kategorienschwelle sehr leicht, die letzte hingegen sehr schwer (Abbildung 2A); starke Zustimmung kam hingegen sehr selten vor bzw. erforderte eine extrem hohe Kompetenz (Personenparameter Theta +4 bis +6). Die Personen der zweiten Klasse (36%) bevorzugten die extremen Antwortoptionen, da die Schwellen der mittleren Kategorien ungeordnet waren (Abbildung 2B); die Wahrscheinlichkeit für extreme Zustimmung respektive Ablehnung war in dieser Gruppe hoch; mittlere Skalenwerte wurden entsprechend selten gewählt, wie die Konzentration der fünf Schwellen im mittleren Skalenbereich zeigt. Die beiden Gruppen unterschieden sich somit hinsichtlich ihrer Antwortstile. Während für die anderen Skalen Personenhomogenität angenommen werden kann, kann der Rückschluss von den Itemantworten auf den visionären Kompetenzgrad nur über die Zuordnung der Person zu einer der beiden Personenklassen erfolgen. Für eine zuverlässige Messung müssten die Parameter aus einem Zweiklassenmodell mit größerer Stichprobe geschätzt werden, um ein günstigeres Verhältnis von Datenpunkten zu geschätzten Modellparametern zu erreichen. Dies unterstreicht die hohen Anforderungen, die an die Entwicklung von Instrumenten zur Kompetenzerfassung angelegt werden müssen.

Tabelle 3 fasst die Skalenkennwerte auf Basis des ordinalen Rasch-Modells zusammen. Die

TABELLE 3 Skalenwerte: Personenparameter, Skalenmittelwert und Reliabilitätsschätzungen

| | Personenparameter* | | Skalenmittelwert | | Rasch-Reliabilität | Cronbachs α |
|----------------------------|--------------------|------|------------------|-----|--------------------|--------------------|
| | M | SD | M | SD | | |
| 1. Einfluss und Motivation | 1.96 | 1.45 | 3.94 | .56 | .76 | .79 |
| 2. Lernen | 1.64 | 1.95 | 3.73 | .72 | .83 | .87 |
| 3. Selbstmanagement | 2.34 | 1.75 | 4.02 | .57 | .85 | .87 |
| 4. Visionäre Kompetenz | 1.01 | 1.51 | 3.63 | .73 | .79 | .81 |
| 5. Teamentwicklung | 1.66 | 1.71 | 3.90 | .61 | .84 | .86 |
| 6. Initiative | 1.30 | 1.59 | 3.84 | .64 | .63 | .69 |
| 7. Ethisches Verhalten | 2.23 | 2.24 | 3.89 | .79 | .87 | .92 |
| 8. Förderung von anderen | 1.59 | 2.05 | 3.72 | .77 | .87 | .90 |
| 9. Kommunikation | 1.35 | 1.35 | 3.84 | .60 | .76 | .77 |
| 10. Entscheidungsfindung | 1.31 | 1.55 | 3.72 | .61 | .83 | .84 |
| 11. Veränderung | 1.35 | 1.53 | 3.74 | .61 | .83 | .84 |

Anmerkung: * beruhend auf WLE-Schätzungen, Einklassenlösung des Mixed-Rasch-Modells

durchweg positiven durchschnittlichen Personenparameter zeigen an, dass die Probanden dieser Stichprobe den Items der jeweiligen Skala eher zustimmten, die Items für diese Stichprobe also eher leicht waren. Die Skalen erreichten befriedigende bis gute Rasch-Reliabilitäten [.76-.87] nach Andrich (2005), die leicht unter den Werten der Originalversion liegen [.84-.92]. Bis auf die Kompetenz *Initiative* [.63] differenziert das ALCP zwischen Personen recht gut hinsichtlich ihrer Kompetenzgrade. Die Rasch-Reliabilitäten stimmten gut mit den internen Konsistenzschätzungen (nach KTT) überein.

Konstruktvalidität: Interkorrelationen von Führungskompetenzen

Für eine Einschätzung der Konstruktvalidität wurden Korrelationen zwischen den Kompetenzskalen analysiert (Tabelle 4). Extrem hohe Skaleninterkorrelationen würden die Frage aufwerfen, ob die Führungskompetenzen nicht unökonomisch und jenseits der zuverlässigen Unterscheidbarkeit erfasst werden. Tatsächlich korrelierten sämtliche Kompetenzen signifikant mit positivem Vorzeichen und eher im mittleren Bereich: Die arithmetischen Mittelwerte der Korrelationskoeffizienten jeder einzelnen Skala beliefen sich auf $.38 \leq r \leq .56$ ($ps < .01$). Dabei zeigt das deutsche ALCP ein ähnliches Interkorrelationsmuster wie das amerikanische Original. Die höchsten einzelnen Zusammenhänge fanden sich zwischen Entscheidungsfindung und Veränderung ($r = .74$) sowie Kommunikation ($r = .69$). Numerisch ebenso hoch korrelierten die Skalen Förderung von anderen und Teamentwicklung ($r = .69$). Somit korrelierten gerade jene Kompetenzen hoch, die entweder Aufgaben- oder Personenorientierung erfassen. Die geringsten Bezüge fanden sich für die Korrelationen von Ethischem Verhalten mit Einfluss und Motivation ($r = .31$), Entscheidungsfindung ($r = .33$) sowie Lernen ($r = .28$). Insgesamt stufen wir die vorliegenden Interkorrelationen als akzeptabel ein. Allerdings korrelieren einzelne Skalen durchaus so hoch, dass bei Berücksichtigung von Messfehlern die Frage nach der Redundanz dieser Skalen aufgeworfen werden könnte. Eine Auswahl von ALCP-Kompetenzen nach Firmen- bzw. Positionsanforderungen erscheint sinnvoll und geboten.

Unklar blieb bislang – auch in Arbeiten zur amerikanischen Version – inwieweit die elf Skalen allesamt eine psychologische Verwandtschaft („Globale Führungskompetenz“) erkennen lassen oder ob sie auf latenter Ebene als elf inhaltlich separierbare Führungsdimensionen aufgefasst werden dürfen. Verwendete man alle ALCP-Items als Indikatoren einer einzigen Dimension, würde man die Eindimensionalität sämtlicher 65 Items ungeprüft voraussetzen und dabei eine facettierte Struktur der Kompetenzen negieren. Analysiert man hingegen die Kompetenzen gemeinsam als elf unabhängige, jeweils einzeln eindimensionale Rasch-Modelle, erhält man über die messfehlerbereinigten Korrelationen der Personenparameter Aufschluss über die Skalen-Abhängigkeiten. Daran anschließend lassen sich allerdings anhand der Personenparameterkorrelationen Hypothesen über die *empirische* Anzahl der dem ALCP zugrundeliegenden psychologischen Dimensionen generieren.

In einem nächsten Schritt wurde daher ein mehrdimensionales Partial-Credit-Modell, welches die Rasch-Modell-Konformität sämtlicher elf Skalen

TABELLE 4 Deskriptive Statistiken von Personen mit und ohne Führungsverantwortung (inkl. Effektstärke) sowie Skalen-Interkorrelationen

| | Deskriptive Statistiken | | | | Skalen-Interkorrelationen (N = 311) | | | | | | | | | | | |
|--|-------------------------|-----|--------------------|-----|-------------------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|--|
| | FK (n = 206) | | Non-FK (n = 90) | | | | | | | | | | | | | |
| | M | SD | M | SD | Cohen's d | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | |
| 1. Einfluss und Motivation | 4.04 | .49 | 3.66 | .63 | .67 | .53 | .61 | .50 | .60 | .42 | .31 | .55 | .55 | .51 | .50 | |
| 2. Lernen | 3.79 | .70 | 3.55 | .76 | .33 | | .37 | .55 | .56 | .37 | .28 | .64 | .49 | .54 | .50 | |
| 3. Selbstmanagement | 4.12 | .55 | 3.78 | .59 | .60 | | | .56 | .54 | .49 | .42 | .45 | .50 | .50 | .54 | |
| 4. Visionäre Kompetenz | 3.75 | .73 | 3.39 | .66 | .52 | | | | .51 | .43 | .42 | .53 | .44 | .59 | .63 | |
| 5. Teamentwicklung | 3.98 | .60 | 3.67 | .60 | .52 | | | | | .45 | .51 | .69 | .62 | .58 | .52 | |
| 6. Initiative | 3.90 | .64 | 3.68 | .64 | .34 | | | | | | .41 | .39 | .44 | .46 | .47 | |
| 7. Ethisches Verhalten | 4.02 | .76 | 3.59 | .81 | .52 | | | | | | | .42 | .38 | .33 | .36 | |
| 8. Förderung von anderen | 3.77 | .74 | 3.54 | .84 | .29 | | | | | | | | .63 | .64 | .65 | |
| 9. Kommunikation | 3.92 | .60 | 3.67 | .59 | .42 | | | | | | | | | .69 | .60 | |
| 10. Entscheidungsfindung | 3.78 | .60 | 3.55 | .65 | .37 | | | | | | | | | | .74 | |
| 11. Veränderung | 3.81 | .59 | 3.57 | .64 | .39 | | | | | | | | | | | |
| Anmerkungen: FK = Führungskräfte, Non-FK = ohne Führungsverantwortung; $r > .15$, $p < .01$ | | | | | | | | | | | | | | | | |

Anmerkungen: FK = Führungskräfte, Non-FK = ohne Führungsverantwortung; $r > .15$, $p < .01$.

simultan überprüft, zu Demonstrationszwecken zugrundegelegt (Wu & Adams, 2006). Während herkömmliche KTT-Strukturgleichungsmodelle ebenfalls multidimensionale Skalenanalysen gestatten, erlaubt die IRT darüber hinaus die Testung einzelner Itemfits sowie die Modellierung der beobachteten Antworten. Multidimensionale Rasch-Modelle bewahren die Struktur der Kompetenzen und der darin enthaltenen Items, während das Gesamtmodell geschätzt und die Item- und Personenparameter kalibriert werden (Rost & Carstensen, 2002). Da das ALCP keine Vorgaben im Hinblick auf die Art und den Grad der Kompetenzüberlappungen macht, wurde die ursprüngliche 11-Dimensionalität beibehalten. Es kann aber auch nicht ausgeschlossen werden, dass elf Dimensionen das Kompetenzspektrum überspezifizieren, und somit in weiterer Forschung eine reduzierte Dimensionsanzahl zur Erklärung der Korrelationen zwischen Skalen und Items nahegelegt wird.

Die Software Acer ConQuest 2.0 (Wu et al., 2007) ist in der Lage, die Parameter des 11-dimensionalen Partial-Credit Modells zu schätzen. Als Item-Fit-Indizes werden antwortresiduenbasierte Mean Square (MNSQ)-Werte verwendet. Hierbei werden die Abweichungen der beobachteten Antworten aller Personen von den unter dem Modell erwarteten Antwortwahrscheinlichkeiten quantifiziert, wobei bei optimaler Modellpassung ein MNSQ-Wert von 1.0 erwartet wird. Es wurde ein Modell mit between-item Dimensionalität gewählt (Hartig & Höhler, 2008, 2009). Hierbei werden die Items der elf hypothetischen Skalen jeweils genau einer latenten Dimension zugeordnet, so dass die elf Dimensionen keine überlappenden Items beinhalten. Zugrundegelegt wurde ein Monte-Carlo-Verfahren (mit 1.000 Monte-Carlo-Integrationsknoten), das in 115 Iterationen konvergierte. Da das Verhältnis von Datenpunkten zu Modellparametern in dieser Analyse ungünstig ist, betonen wir ausdrücklich, dass anhand dieser Analysen lediglich exemplarisch aufgezeigt werden soll, welchen Mehrwert eine solche Modellierung haben kann. Für Messungen von Kompetenzen muss ein psychometrischer Gütenachweis erbracht werden, und dafür sind im Sinne einer validen Rasch-Modellierung viele Datenpunkte nötig (vgl. Stone & Yumoto, 2004).

Voraussetzung für die Interpretation der Skalen-Interkorrelationen ist die Modellkonformität jeder einzelnen Skala. Lediglich zwei Items aus der Skala *Einfluss* und *Motivation* sowie ein Item der Skala *Teamentwicklung* wiesen mit gewichteten MNSQ-Statistiken von 1.32 ($T = 3.6$), 1.50 ($T = 5.3$) und 1.18 ($T = 2.1$) einen signifikanten Underfit auf, wobei der erste und der letzte Wert noch vernachlässigbare Abweichungen darstellen, wenn man die MNSQ-Werte als Effektstärken mit einem Toleranzbereich von $MNSQ = 0.8$ -1.2 oder gar 0.75-1.33 interpretiert (vgl. Wilson, Allen & Li, 2006). Das bedeutet, dass lediglich ein einziges dieser drei Items durch die Annahme der Eindimensionalität der zugrundeliegenden Skala unzureichend beschrieben wird. Angesichts der Größenordnung ist selbst im Fall dieses einen Items die Qualität der Kompetenzmessung noch nicht gefährdet. Daneben wiesen 6 Items einen Overfit auf ($MNSQ = .76$ -.83, $T = -3.1$ bis -2.1). Aus Sicht der Modellspezifikation ist Overfit suboptimal und deutet auf Determination (z.B. durch Itemredundanzen) hin, jedoch ist dies weniger schwerwiegend als Underfit, bei dem ein Item die Merkmalsdimension nicht ausreichend abbildet (Engelhard & Stone, 1998).

Tabelle 5 zeigt die messfehlerbereinigten Zusammenhänge zwischen den elf Kompetenzen. Das Befundmuster ähnelt dem der manifesten Korrelationen (s.o.). Insgesamt weisen sämtliche Kompetenzen die erwarteten positiven Korrelationen auf, jedoch bilden nicht sämtliche Skalen zusammen nur eine einzige globale Führungsdimension ab ($r_s < 1$). Die starken statistischen Assoziationen zwischen den Kompetenzen dürfen jedoch nicht dahingehend interpretiert werden, dass eine Zusammen-

TABELLE 5 Varianzen, Kovarianzen (oberhalb der Diagonale) und latente Korrelationen (unterhalb der Diagonale) der ALCP-Personenparameter

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|----------------------------------|------|------|------|------|------|------|------|------|------|------|------|
| 1 Einfluss und Motivation | • | 1.10 | 1.31 | 1.07 | 1.52 | 0.84 | 1.46 | 1.29 | 1.28 | 0.93 | 1.04 |
| 2 Lernen | .77 | • | 0.68 | 1.06 | 1.20 | 0.66 | 1.06 | 1.26 | 1.07 | 0.89 | 1.01 |
| 3 Selbstmanagement | .86 | .45 | • | 0.98 | 1.23 | 0.81 | 1.32 | 0.93 | 1.11 | 0.84 | 1.01 |
| 4 Visionäre Kompetenz | .84 | .85 | .74 | • | 1.19 | 0.77 | 1.40 | 1.18 | 1.10 | 0.95 | 1.29 |
| 5 Teamentwicklung | .91 | .74 | .71 | .82 | • | 1.03 | 2.21 | 1.79 | 1.55 | 1.13 | 1.25 |
| 6 Initiative | .85 | .69 | .79 | .90 | .91 | • | 1.45 | 0.96 | 0.87 | 0.68 | 0.79 |
| 7 Ethisches Verhalten | .54 | .41 | .47 | .60 | .72 | .80 | • | 2.05 | 1.63 | 1.29 | 1.30 |
| 8 Förderung von anderen | .73 | .73 | .50 | .77 | .89 | .81 | .63 | • | 1.59 | 1.23 | 1.43 |
| 9 Kommunikation | .89 | .76 | .74 | .87 | .94 | .98 | .62 | .91 | • | 1.10 | 1.26 |
| 10 Entscheidungsfindung | .79 | .78 | .68 | .94 | .84 | .86 | .60 | .87 | .95 | • | 1.10 |
| 11 Veränderung | .74 | .73 | .69 | .92 | .78 | .83 | .50 | .84 | .91 | .98 | • |
| <i>Varianz</i> | 1.47 | 1.39 | 1.59 | 1.11 | 1.91 | 0.67 | 4.94 | 2.13 | 1.42 | 0.94 | 1.36 |

legung von Skalen zwangsläufig geboten ist. Eine explorative Faktorenanalyse der Items zeigt, dass mindestens acht Faktoren mit Eigenwerten größer eins zur Aufklärung von mehr als 50% der Varianz benötigt werden. Während der Scree-Plot nach sechs Faktoren tendenziell abflacht, sind laut MAP-Verfahren (minimum average partial correlation method; Velicer, 1976; Velicer, Eaton & Fava, 2000) mindestens neun Hauptkomponenten zur Aufklärung der vorhandenen Item-Interkorrelationen erforderlich. Die vorgefundene Skalen-Redundanz erlaubt aber, ohne größeren Informationsverlust nur solche Kompetenzen zu erheben, die situationsspezifisch oder unter Akzeptanz Gesichtspunkten besonders passen.

Inhalts- bzw. Kriteriumsvalidität: Diskrimination bekannter Gruppen

Neben den Itemwortlauten sollte der von Kompetenzdefinitionen geforderte Erfolgsaspekt empirisch nachweisbar sein (Catano, 1998). In einer ersten Annäherung überprüften wir, ob Unterschiede zwischen bekannten Gruppen identifizierbar sind. Dabei orientierten wir uns am Vorgehen von Popper und Kollegen (2004), die Gruppenvergleiche zwischen Führungskräften und Nicht-Führungskräften durchführten. Unsere Testpersonen in Positionen mit aktueller Personalverantwortung ($n = 206$) wurden mit jenen ohne leitende Funktion ($n = 90$) verglichen: Die Ergebnisse einer MANCOVA mit Alter und Geschlecht als Kovariaten zeigten über alle Skalen hinweg unterschiedliche Kompetenzgrade, Pillai-Spur- $F(11, 282) = 3.78, p < .01, \eta^2 = .13$ (Tabelle 4). Ein Vergleich von adjustierten und nicht-adjustierten Skalenmittelwerten ergab – in Übereinstimmung mit den Ergebnissen aus dem Mixed-Rasch-Modell – einen vernachlässigbaren Einfluss von Alter und Geschlecht auf die Kompetenzmessungen. Die Kompetenzmessung kann damit als unabhängig von diesen potentiellen Moderatoren gelten und bedarf keiner alters- oder geschlechtsspezifischen Normen. Es fanden sich durchgängig höhere Kompetenzen bei Führungskräften, $ts(294) > 2.36, ps < .02$, Cohen's $ds = 0.29-0.67$. Die größten Differenzen traten bei den Kompetenzen *Einfluss* und *Motivation* sowie *Selbstmanagement* auf. Auf den Skalen *Lernen*, *Initiative*, *Förderung von anderen*, *Entscheidungsfindung* und *Veränderung* sind die Kompetenzunterschiede substantiell, aber eher gering. Wie für Kompetenzen gefordert, differenziert das ALCP damit konsistent zwischen führungserfahrenen und -unerfahrenen Personen, wenngleich unklar ist, ob Führungskräfte auf allen Dimensionen (z.B. ethisches Verhalten) besser abschneiden müssen (vgl. Hoyt, Price & Poatsy, 2013).

Diskussion

Unsere Studie zeigt die probabilistisch-testtheoretische Analyse des ALCP nach dem ordinalen Rasch-Modell auf, welches ein 11-dimensionales Kompetenzmodell annimmt. Der Vorteil des IRT-Ansatzes liegt in einer kritischen Überprüfung dreier Annahmen, die üblicherweise explizit (oder auch nur implizit) an Fragebögen gestellt werden:

- a) dass alle Items einer Skala dasselbe Personmerkmal erfassen (Itemhomogenität),
- b) dass alle Personen die Skalen aufgrund desselben Personmerkmals beantworten (Personenhomogenität), und
- c) dass Antwortskalen adäquat (d.h. in Abhängigkeit von der Merkmalsausprägung) verwendet werden und somit nicht über- oder unterdifferenziert sind.

Wir gingen der psychometrischen Qualität des ALCP und ersten Validitätshinweisen im deutschsprachigen Raum nach. Die meisten der Items und Skalen erfüllten sämtliche Kriterien im Rahmen des ordinalen Rasch-Modells: Itemhomogenität, Personenhomogenität und geordnete Schwellenparameter. Insgesamt sprechen die Daten dafür, dass die Annahme der Rasch-Homogenität der ALCP-Führungskompetenzen wie im amerikanischen Original beibehalten werden kann. Die Kompetenzen bilden in zuverlässiger Weise erfolgsrelevante Aspekte ab. Unter praktischen Gesichtspunkten erlaubt diese Modellkonformität dem Diagnostiker die modulare Auswahl der in einem Unternehmen oder in einer Beurteilungssituation als relevant erachteten Kompetenzen. Beispielsweise korrelierten sowohl die messfehlerbereinigten Personenparameter als auch die unkorrigierten Skalenwerte von *Entscheidungsfindung* und *Veränderung* bei uns so hoch, dass im Unternehmen die Auswahl des passenderen Merkmals aus ökonomischen Gründen gerechtfertigt wäre (vgl. Tabellen 4 und 5).

Die Analysen der deutschen Items bestätigten die positiven psychometrischen Eigenschaften des amerikanischen Originals für die Mehrzahl der Kompetenzen. Fast keines der 65 ins Deutsche übersetzten Items musste als modell-nonkonform verworfen werden. Soweit geringfügige Schwellenparametervertauschungen vorlagen, konnten diese auf die geringe Wahl einzelner Antwortkategorien zurückgeführt werden. Die Skalen erreichten bis auf eine Skala befriedigende bis gute Rasch-Reliabilitäten.

Die ALCP-Items besitzen gute psychometrische Eigenschaften, was auf ausreichend zuverlässige und homogene Kompetenzmessungen hindeutet. Für die Mehrzahl der Skalen ist die Einklassenlösung zu favorisieren. Einzig bei *Visionäre Kompetenz* sprachen die Ergebnisse für eine Zweiklassenlösung, die als Personen mit Tendenz zur Mitte und solche mit extremem Antwortverhalten identifiziert wurden. Bei dieser Skala kann die Personmerkmalsausprägung deshalb nicht unabhängig von der Klassenzugehörigkeit als Kompetenz interpretiert werden und ist somit psychometrisch problematisch.

Die Überprüfung der ALCP-Skalen- und Itemeigenschaften nach dem Rasch-Modell ist für Selbst- wie Fremdeinschätzung (360°-Feedback) von großer Bedeutung. Dank geordneter Antwortkategorien erlauben Rasch-Modelle linear interpretierbare Kompetenzeinstufungen (Wright & Masters, 1982), wodurch interindividuelle Kompetenzunterschiede gemessen werden können. Es handelt sich demonstrierbar um Niveauunterschiede in Kompetenzen und nicht lediglich um Unterschiede im Antwortverhalten im Vergleich zu einer irgendwie gearteten Normierungsstichprobe.

Für die Validität des deutschsprachigen ALCP sprechen auch die weiteren Ergebnisse. Führungskräfte unterschieden sich in ihrer selbstberichteten Kompetenz nachweislich von Personen ohne aktuelle Führungstätigkeiten. Abweichend von Popper und Kollegen (2004) zeigt unser Gruppenvergleich, dass Unterschiede zwischen Führungskräften und Nicht-Führungskräften eben nicht nur für allgemeine Persönlichkeitskonstrukte (wie Ängstlichkeit und Optimismus), sondern auch für spezifisch führungsrelevante Merkmale nachweisbar sind. Da sie sich ausschließlich auf Führungstätigkeiten beziehen, sind ALCP-Führungskompetenzen potentiell sparsamer – und zwar ohne Bedeutungsüberschuss – als die breiteren Persönlichkeitsmerkmale. Die Tatsache, dass das ALCP klar zwischen Personen mit Führungsaufgaben und solchen ohne trennen konnte, ist umso beachtenswerter, als die beiden untersuchten Gruppen in sich auch sehr heterogen waren und aus verschiedenen Branchen rekrutiert wurden. Es ist davon auszugehen, dass die Unterschiede innerhalb eines einzelnen Unternehmens, in der die Gruppe der Führungskräfte und die Gruppe der Mitarbeiter ohne Leitungsfunktion in sich homogener sind, noch deutlicher ausfallen dürften.

Jenseits der Frage nach psychometrischen Gütekriterien scheint der Bezug zu mehreren Führungstheorien zunächst wenig ökonomisch. Ebenso scheint das Fehlen weiterer beruflicher Verhaltensweisen – neben Führungsaspekten – ein Nachteil zu sein. Führungskompetenzmodelle stellen jedoch letztlich den Versuch dar, Überlappungen und nebeneinander unterschiedlicher Führungstheorien zu überwinden und gleichzeitig den kontextabhängigen, aber praxisrelevanten Ansprüchen an Führungskräfte durch Auswahl anforderungsanalytisch wichtiger Kompetenzen gerecht zu werden. Optimistisch stimmt daher, dass die Rasch-Homogenität – mit Einschränkungen bei der Skala *Visionäre Kompetenz* – kultur- und sprachübergreifend ohne größere Modifikationen repliziert werden konnte. Sollte die hier vorgestellte psychometrische Entwicklungslinie Einzug in die Testkonstruktion bei anderen Kompetenzmodellen finden, könnten Kompetenzen künftig qualitativ hochwertig im Selbstbericht erfasst werden. Hierbei sollten wie beim ALCP die folgenden Schritte eingehalten werden:

- 1) Erarbeitung eines möglichst großen Itempools zur Erfassung relevanter Kompetenzen;
- 2) Vorgabe der Items an eine möglichst große Stichprobe;
- 3) Überprüfung der Item- und Skalenqualität mittels eindimensionalem Rasch-Modell;
- 4) Auswahl der psychometrisch guten Items und Skalen;
- 5) Überprüfung des Fragebogens auf Personenhomogenität (z.B. mit dem mixed Rasch-Modell), so dass sichergestellt ist, dass Unterschiede zwischen Gruppen auf tatsächliche Unterschiede in den Kompetenzen zurückzuführen sind.

Zu guter Letzt sei auch daran erinnert, dass mit IRT-basierten Verfahren nicht nur die Chance auf unternehmensübergreifenden Einsatz psychometrisch fundierter Verfahren zur Erfassung von Kompetenzen wächst, wobei spezifische Normstichproben entfallen, sondern dass mit fremdsprachlicher Verbreitung desselben Kompetenzgefüges die Attraktivität für international operierende Firmen steigt. Sofern die Modellannahmen wie beim ALCP gelten, kann (bei korrekter Übersetzung) der Einsatz auch in anderen Ländern angestrebt werden.

Einschränkung und Ausblick

Das hier vorgestellte Vorgehen der probabilistisch-testtheoretischen Fundierung der Messinstrumente zu Kompetenzmodellen hat Beispielcharakter, doch eine größere Stichprobe würde belastbarere Resultate unserer Analysen ermöglichen. Unser Vorgehen bietet sich für jegliche Entwicklung von (firmen-)spezifischen Kompetenzmodellen an und erlaubt die Überprüfung von Modellannahmen über unterschiedliche Gruppen und Kontexte hinweg, auch jenseits von Führungsaspekten.

Zukünftige Studien sollten die Konstrukt- und kriteriumsbezogene Validität differenzierter beleuchten. Vor einem uneingeschränkten Einsatz im Unternehmen würde der Psychometriker zunächst eine Wunschliste zum ACLP formulieren. Auf dieser Liste würden die folgenden Aspekte weit oben rangieren: Zum einen sollten an einer noch größeren und ausreichend heterogenen Stichprobe von etablierten und angehenden Führungskräften erneut die Skaleninterkorrelationen im Sinne der Redundanzprüfung inspiziert werden. Wie üblich bei reinen Selbstberichtsdaten kann nicht ausgeschlossen werden, dass Führungskräfte lediglich stärker im Sinne des Stereotyps, das auf sie angewendet wird, geantwortet haben. Um die potentielle Inflation der Korrelationen durch dieselben Rater (Methodenvarianz, same-source bias) abschätzen zu können, sollten also Fremdeinschätzungen zu sämtlichen Kompetenzen einer Führungskraft das Bild komplettieren. Interessant wären auch nicht nur die Zusammenhänge der ALCP-Skalen untereinander, sondern inwieweit sich Bezüge zu etablierten Persönlichkeitsmerkmalen herstellen lassen, z.B. zu den klassischen Big Five Traits wie im NEO-PI-R (Ostendorf & Angleitner, 2004) oder den Skalen des Bochumer Inventars zur berufsbezogenen Persönlichkeitsbeschreibung (Hossiep & Paschen, 2003).

Möglicherweise könnten die Kompetenzen auch weitgehend durch Persönlichkeitseigenschaften aufgeklärt werden. Interventionsstudien könnten wiederum zeigen, dass ein Training ausgewählte Kompetenzen gezielt verbessern kann, während andere sich nicht ändern. Hinsichtlich der Kriteriumsvalidität wären weitere Berufserfolgskriterien jenseits von Selbstauskünften wünschenswert, z.B. objektive Gehaltsermittlung, Vorgesetztenbeurteilungen und Mitarbeiterevaluationen der Leistungsfähigkeit der Führungskräfte.

Literatur

- Andrich, D. (2005). Rasch models for ordered response categories. In B. Everitt & D. Howell (Hrsg.), *Encyclopedia of Statistics in Behavioral Science* (Vol. 4, S. 1698-1707). New York: John Wiley & Sons.
- Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Brannick, M. T. & Levine, T.L. (2002). *Job analysis: Methods, research, and applications for human resource management in the new millennium*. Thousands Oaks: Sage.
- Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion*. München: Pearson.
- Catano, V. M. (1998). *Competencies: A review of the literature and bibliography (CCHRA-PHASE I Report)*. Halifax, Nova Scotia: Saint Mary's University.
- Davier, M. von (2001). *WINMIRA 2001*. Zugriff am 20.07.2008 <http://winmira.von-davier.de>
- Engelhard, G. & Stone, G. E. (1998). Evaluating the quality of ratings obtained from standard-setting judges. *Educational and Psychological Measurement*, 58, 179-196.
- Erpenbeck, J. & Rosenstiel, L. v. (Hrsg.) (2003). *Handbuch Kompetenzmessung. Erkennen, verstehen und bewerten von Kompetenzen in der betrieblichen, pädagogischen und psychologischen Praxis*. Stuttgart: Schäffer-Poeschel.
- Felfe, J. (2006). Validierung einer deutschen Version des „Multifactor Leadership Questionnaire“ (MLQ Form 5x Short) von Bass und Avolio (1995). *Zeitschrift für Arbeits- und Organisationspsychologie*, 50, 61-78.
- Hartig, J. & Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within-item and between-item multidimensionality. *Zeitschrift für Psychologie/Journal of Psychology*, 216, 89-101.
- Hartig, J. & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies In Educational Evaluation*, 35, 57-63.
- Harris, M. (1998). Competency modeling: Viagrazed job analysis or impotent imposter? *The Industrial-Organizational Psychologist*, 36, 37-41.
- Hossiep, R. & Paschen, M. (2003). *Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung – BIP* (2. Aufl.; unter Mitarbeit von O. Mühlhaus). Göttingen: Hogrefe.
- Hoyt, C. L., Price, T. L. & Poatsy, L. (2013). The social role theory of unethical leadership. *Leadership Quarterly*, 24, 712-723.
- Laber, M. E. & O'Connor, J.K. (2000). Competency modeling: ready, set, research. *The Industrial-Organizational Psychologist*, 37, 91-96.
- Maij-de Meij, A. M., Kelderman, H. & van der Flier, H. (2008). Fitting a mixture item response theory model to personality questionnaire data. *Applied Psychological Measurement*, 32, 611-631.
- Markus, L. H., Cooper-Thomas, H. D. & Allpress, K. N. (2005). Confounded by competencies? An evaluation of the evolution and use of competency models. *New Zealand Journal of Psychology*, 34, 117-126.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.
- McClelland, D. C. (1973). Testing for competence rather than intelligence. *American Psychologist*, 28, 1-14.
- Ostendorf, F. & Angleitner, A. (2004). *NEO-Persönlichkeitsinventar nach Costa und McCrae*. Revidierte Fassung. Göttingen: Hogrefe.
- Popper, M., Amit, K., Gal, R. et al. (2004). The capacity to lead: Major psychological differences between leaders and nonleaders. *Military Psychology*, 16, 245-263.
- Rosenstiel, L. v. (2003). Grundlagen der Führung. In L. v. Rosenstiel, E. Regnet & M. E. Domsch (Hrsg.), *Führung von Mitarbeitern* (5., überarb. Aufl., S. 3-25). Stuttgart: Schäffer-Poeschel.
- Rost, J. (1990). Rasch Models in latent classes. An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14, 271-282.
- Rost, J. (2004). *Lehrbuch Testtheorie Testkonstruktion*. Bern: Huber.
- Rost, J. & Carstensen, C. H. (2002). Multidimensional Rasch measurement via item component models and faceted designs. *Applied Psychological Measurement*, 26, 42-56.
- Rost, J. & von Davier, M. (1994). A conditional item fit index for Rasch models. *Applied Psychological Measurement*, 18, 171-182.
- Schafer, J. L. & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147-177.
- Schuler, H. (2001). Arbeits- und Anforderungsanalyse. In H. Schuler (Hrsg.), *Lehrbuch der Personalpsychologie* (S. 43-61). Göttingen: Hogrefe.
- Sherron, C. T. (2000). *Psychometric development of the Adaptive Leadership Competence Profile*. Elektronisch veröffentlichte Dissertation, University of North Texas: <http://digital.library.unt.edu/permalink/meta-dc-2697>
- Schippmann, J. S., Ash, R. A., Battista, M., Carr, L., Eyde, L. D., Hesketh, B. et al. (2000). The practice of competency modeling. *Personnel Psychology*, 53, 703-740.
- Smith, E. V. jr. (2005). Effect of item redundancy on Rasch item and person estimates. *Journal of Applied Measurement*, 6, 147-163.
- Spencer, L. M., McClelland, D. C. & Spencer, S. M. (1990). *Competency assessment methods: History and state of the art*. Paper presented at the American Psy-

- chological Association Annual Conference, Boston, Mass., August 13, 1990.
- Stogdill, R. M. (1963). *Manual for the LBDQ Form XII*. Columbus: Ohio State University.
- Stone, M. & Yumoto, F. (2004). The effect of sample size for estimating Rasch/IRT parameters with dichotomous items. *Journal of Applied Measurement*, 5, 48-61.
- Testkuratorium der Föderation deutscher Psychologenverbände (1986). Beschreibung der einzelnen Kriterien für die Testbeurteilung. *Diagnostica*, 32, 358-360.
- Tett, R. P., Jackson, D. N., Rothstein, M. & Reddon, J. R. (1999). Meta-analysis of bidirectional relations in personality – job performance research. *Human Performance*, 12, 1-29.
- Tuerlinckx, F. & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods*, 6, 181-195.
- Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41, 321-327.
- Velicer, W. F., Eaton, C. A. & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin & E. Helmes (Hrsg.), *Problems and solutions in human assessment: Honoring Douglas Jackson at seventy* (S. 41-71). Boston: Kluwer.
- Wilson, M., Allen, D. D. & Li, J. C. (2006). Improving the measurement in health education and health behavior research using item response modeling: Introducing item response modeling. *Health Education Research*, 21 (Supplement 1): i4–i18.
- Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: Mesa Press.
- Wu, M.L. & Adams, R.J. (2006). Modelling mathematics problem solving item responses using a multidimensional IRT model. *Mathematics Education Research Journal*, 18, 93-113.
- Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S.A. (2007). *ACER ConQuest Version 2: Generalised item response modelling software* [computer program]. Camberwell: Australian Council for Educational Research.



PROF. DR. RICARDA
STEINMAYR
Institut für Psychologie
FK 12 Erziehungswissenschaft
und Soziologie
TU Dortmund
Emil-Figge-Straße 50
D-44227 Dortmund
ricarda.steinmayr@tu-
dortmund.de



DR. MATTHIAS BLUEMKE
Universität Heidelberg



PROF. DR. MORITZ HEENE
Ludwig-Maximilians-
Universität München



DR. TANJA BIPP
Eindhoven University of
Technology and Open
University of the Netherlands